Scheduling and Control of Queueing Networks

Applications of queueing network models have multiplied in the last generation, including scheduling of large manufacturing systems, control of patient flow in health systems, load balancing in cloud computing, and matching in ride sharing. These problems are too large and complex for exact solution, but their scale allows approximation.

This book is the first comprehensive treatment of fluid scaling, diffusion scaling, and many-server scaling in a single text presented at a level suitable for graduate students. Fluid scaling is used to verify stability, in particular treating max weight policies, and to study optimal control of transient queueing networks. Diffusion scaling is used to control systems in balanced heavy traffic, by solving for optimal scheduling, admission control, and routing in Brownian networks. Many-server scaling is studied in the quality and efficiency driven Halfin–Whitt regime and applied to load balancing in the supermarket model and to bipartite matching in ride-sharing applications.

GIDEON WEISS is Professor Emeritus in the Department of Statistics at the University of Haifa, Israel. He has previously held tenured positions at Tel Aviv University and at Georgia Tech Industrial and Systems Engineering and visiting positions at Berkeley, MIT, Stanford, NYU, and NUS. He is author of some 90 research papers and served on the editorial boards of leading journals on operations research and applied probability. His work includes significant contributions to the fields of time series, stochastic scheduling, bandit problems, fluid analysis of queueing networks, continuous linear programming, and matching problems.

INSTITUTE OF MATHEMATICAL STATISTICS TEXTBOOKS

IMS Textbooks give introductory accounts of topics of current concern suitable for advanced courses at master's level, for doctoral students and for individual study. They are typically shorter than a fully developed textbook, often arising from material created for a topical course. Lengths of 100–290 pages are envisaged. The books typically contain exercises.

In collaboration with the International Society for Bayesian Analysis (ISBA), selected volumes in the IMS Textbooks series carry the "with ISBA" designation at the recommendation of the ISBA editorial representative.

Other Books in the Series (*with ISBA)

1. *Probability on Graphs,* by Geoffrey Grimmett
2. *Stochastic Networks,* by Frank Kelly and Elena Yudovina
3. *Bayesian Filtering and Smoothing,* by Simo Särkkä
4. *The Surprising Mathematics of Longest Increasing Subsequences,* by Dan Romik
5. *Noise Sensitivity of Boolean Functions and Percolation,* by Christophe Garban and Jeffrey E. Steif
6. *Core Statistics,* by Simon N. Wood
7. *Lectures on the Poisson Process,* by Günter Last and Mathew Penrose
8. *Probability on Graphs (Second Edition),* by Geoffrey Grimmett
9. *Introduction to Malliavin Calculus,* by David Nualart and Eulàlia Nualart
10. *Applied Stochastic Differential Equations,* by Simo Särkkä and Arno Solin
11. **Computational Bayesian Statistics,* by M. Antónia Amaral Turkman, Carlos Daniel Paulino, and Peter Müller
12. *Statistical Modelling by Exponential Families,* by Rolf Sundberg
13. *Two-Dimensional Random Walk: From Path Counting to Random Interlacements,* by Serguei Popov

Scheduling and Control of Queueing Networks

GIDEON WEISS
University of Haifa, Israel

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108415323
DOI: 10.1017/9781108233217

First published 2022

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Weiss, Gideon, author.
Title: Scheduling and control of queueing networks / Gideon Weiss,
The University of Haifa.
Description: First edition. | New York : Cambridge University Press, [2021] |
Series: Institute of mathematical statistics textbooks | Includes
bibliographical references and index.
Identifiers: LCCN 2021029970 | ISBN 9781108415323 (hardback)
Subjects: LCSH: Queuing theory. | System analysis–Mathematics. |
Scheduling–Mathematics. | BISAC: MATHEMATICS / Probability & Statistics / General
Classification: LCC QA274.8 .W43 2021 | DDC 519.8/2–dc23
LC record available at https://lccn.loc.gov/2021029970

ISBN 978-1-108-41532-3 Hardback
ISBN 978-1-108-40117-3 Paperback

Additional resources for this publication at www.cambridge.org/9781108415323

Contents

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Frontmatter
More Information

Cambridge University Press
978-1-108-41532-3 — Scheduling and Control of Queueing Networks
Gideon Weiss
Frontmatter
More Information

Notation

$\mathbb{P}$ probability
$\mathbb{E}$ expectation
$\mathbb{1}$ characteristic function
$\mathbb{P}_x$ probability from initial state x
$\mathbb{E}_x$ expectation from initial state x
$\mathbb{N}$ natural numbers, $0, 1, \ldots$
$\mathbb{Z}$ integers
$\mathbb{R}$ the real line
$\mathbb{C}$ space of continuous functions
$\mathbb{D}$ space of functions right continuous with left limits
$\mathbf{1}$ vector of 1's
π stationary probabilities
$\mathcal{A}(t)$ arrival process
$\mathcal{D}(t)$ departure process
$\mathcal{Q}(t)$ queue length
$\mathcal{W}(t)$ workload process
$\mathcal{S}(t)$ service process
$\mathcal{T}(t)$ busy time
$\mathcal{I}(t)$ idle time
$\mathcal{J}(t)$ free time
a_ℓ arrival time of customer ℓ
T_ℓ interarrival time, $T_\ell = a_\ell - a_{\ell-1}$ (Chapters 1–4)
u_ℓ interarrival time, $u_\ell = a_\ell - a_{\ell-1}$ (Chapters 5–20)
X_ℓ service requirement of customer ℓ (Chapters 1–4)
υ_ℓ service requirement of customer ℓ (Chapters 5–20)
F distribution of interarrival times
G distribution of service times
H distribution of patience times
V_ℓ waiting time of customer ℓ
$\bar{V}$ mean waiting time
W_ℓ sojourn time of customer ℓ
$\bar{W}$ mean sojourn time
$\bar{\mathcal{W}}$ mean workload
ρ offered load or traffic intensity
α_i exogenous arrival rates
λ_i total arrival rates
μ_i service rates
$p_{i,j}$ routing probabilities
ν_k nominal allocation
c_a coefficient of variation of interarrival times
c_s coefficient of variation of service time
C constituency matrix
C_i constituency of server i
A resource consumption matrix
R input-output matrix
R^{-1} work requirement matrix
B_κ a compact neighborhood of the origin
$\mathscr{C}$ set or subset of customer types (Chapter 22)
$\mathscr{S}$ set or subset of server types (Chapter 22)

$\mathcal{U}$ set or subset of customer types unique to some servers (Chapter 22)
$\mathcal{G}$ compatibility graph (Chapter 22)
$\mathcal{P}(J)$ permutations of the set J
$\emptyset$ the empty set
r.h.s. right-hand side
i.i.d. independent identically distributed
u.o.c. uniformly on compacts
a.s. almost surely
c.o.v. coefficient of variation
pdf probability density function
cdf cumulative distribution function
BM Brownian motion
RBM reflected Brownian motion
RCLL right continuous with left limits
MCQN multi-class queueing networks
PASTA Poisson arrivals see time averages
BP busy period
EFSBP exceptional first service busy period
HOL head of the line policy
FCFS first come first served
LCFS last come first served, preemptive
PS processor sharing
FBFS first buffer first served, in re-entrant line
LBFS last buffer first served, in re-entrant line
SPT shortest processing time
SEPT shortest expected processing time
SRPT shortest remaining processing time
IVQ infinite virtual queue
BCP Brownian control problem
ED efficiency driven service
QD quality driven service
QED quality and efficiency driven service
CRP complete resource pooling
PSBS parallel skilled based service
ALIS assign longest idle server
SD server dependent service rates
QIR queue and idleness ratio policy

Introduction

Queueing networks are all pervasive; they occur in service, manufacturing, communication, computing, internet and transportation. Much of queueing theory is aimed at performance evaluation of stochastic systems. Extending the methods of deterministic optimization to stochastic models so as to achieve both performance evaluation and control is an important and notoriously hard area of research. In this book our aim is to familiarize the reader with recent techniques for scheduling and control of queueing networks, with emphasis on both evaluation and optimization.

Queueing networks of interest are discrete, stochastic dynamical systems, often of very large size, and exact analysis is usually out of the question. Furthermore, the data necessary for exact analysis is rarely available. Thus, to obtain useful results we are led to use approximations. In this book, our emphasis is on approximations obtained from scaled versions of the systems, and analyzing the limiting behavior when the scale tends to infinity. We will be studying three types of scaling: fluid, diffusion, and many server.

Fluid scaling (Part IV): We count time in units of n and space, expressed by number of items, in units of n. This will be a reasonable model to ask what happens to a system with n items in a time span in which n items are processed. Under fluid scaling, a discrete stochastic system may converge to a deterministic continuous process, its fluid model. Fluid models are used in two ways: first, to answer the question of stability – is the system capable of recovering from extreme situations, in which case it may converge to a stationary behavior. Second, perhaps more exciting, we can use fluid scaling to obtain asymptotically optimal control of transient systems over finite time horizons.

Diffusion scaling (Part V): Space is scaled by units of n, and time by units of n^2. On this scale a stable system may reach stationary behavior, and reveal the congested elements of the network in balanced heavy traffic. Approximation of these by stochastic diffusion processes can be used to evaluate performance measures. Furthermore, on the diffusion scale we may formulate and solve Brownian control problems and derive efficient policies.

Many-server scaling (Part VI): Increasingly, recent applications involve systems with many servers and a large volume of traffic. For such systems time is not scaled, but the number of servers and the arrival rates are scaled. These models preserve not just first moment parameters of the fluid scaling and second moment parameters of the diffusion scaling, but the full service time distribution of individual items moving through the system. Many-server models are used to answer staffing-level questions, and to achieve quality of service goals.

The first three parts of the book cover more conventional material, as well as introducing some of the techniques used later. *Part I* covers birth and death queues and the M/G/1 queue, and a chapter on scheduling. *Part II* deals with approximations to G/G/1, introducing fluid and diffusion scaling and many-server G/G/∞. Two chapters, Chapter 5 and Chapter 7 survey some of the essential probability theory background, at a semi-precise level. *Part III* of the book introduces Jackson networks and related queueing networks with product-form stationary distributions, and generalized Jackson networks.

The book is aimed at graduate students in the areas of Operations Research, Operations Management, Computer Science, Electrical Engineering, and Mathematics, with some background in applied probability. It can be taught as a two part course, using the first three parts of the book as a basic queueing course, or it can be taught to students already familiar with queueing theory, where the first three parts are skimmed and the emphasis is on the last three parts. I have taught this material three times, at three different schools, as a PhD-level course in a single semester, though it was somewhat tight to include all of the material of the second half of the book. I tried to make each chapter as self-contained as possible, to enable more flexibility in teaching a course, and to be more useful for practitioners.

Each chapter in the book is followed by a list of sources, and by exercises. Some of the exercises lead to substantial extensions of the material, and I provide references for those. A few problems that require further study and much more effort are in addition marked by (∗). A solution manual will accompany the book, and be placed on the book website, `www.cambridge.org/9781108415323`.

I conclude this introduction with six examples of potential applications that the techniques developed in this book are designed for:

Semiconductor wafer fabrication plant: Given the current state of the plant, how to schedule production for the next six weeks. While this appears at first to be a deterministic job-shop scheduling problem, optimal schedules never work due to unexpected stochastic interference. In Chapter 14 we formulate this as a control problem of a discrete stochastic transient queueing network. We use fluid scaling to obtain and solve a deterministic continuous control problem, and then track the optimal fluid solution using decentralized control.

Input queued crossbar switches: Scheduling the traffic through ultra-high speed communication switches so as to achieve maximum throughput is solved by a maximum pressure policy as described in Chapter 12.

Joint management of operating theaters in a hospital: To control the flow of patients, surgeons, equipment, and theaters on a long-term basis, this can be formulated as a multi-class queueing network, which is operating in stationary balanced heavy traffic. In Chapters 16 and 17 we use diffusion scaling to formulate and solve such problems as stochastic Brownian control problems, and the optimal solution of the Brownian problem is used to determine policies that use admission limits, choice of routes, scheduling priorities, and thresholds.

Control of a call center: This is modeled as a parallel service system, where types of customers are routed to pools of compatible skilled servers. Here design of the compatibility graph, balancing staffing levels, and maintaining acceptable levels of abandonment need to be determined, based on many-server scaling in Chapters 19 and 20.

Cloud computing and web searching: Balancing the utilization of the servers and controlling the lengths of queues at many servers needs to be achieved with a minimum amount of communication. Asymptotic optimality here is achieved by routing tasks to the shorter of several randomly chosen servers. This is modeled by the so-called supermarket problem, as studied in Chapter 21.

Ride sharing: Drivers as well as passengers become available in a random arrival stream, and have limited patience waiting for a match to determine a confirmed trip. Matching available drivers to passengers according to their compatibility, and dispatching on first come first served is analyzed and used to design regimes of operation in Chapter 22.

In writing this book I benefitted from the help of many colleagues, foremost I wish to thank to Ivo Adan, Onno Boxma, Asaf Cohen, Liron Ravner, Shuangchi He, Rhonda Righter, Dick Serfozo, and Hanqin Zhang for their many useful comments and suggestions, and to my students who kept me in check.